

# Stratus Consulting

## **External Review of the Economic Value of Current and Improved Weather Forecasts Study** Final Report

*Prepared for:*

Dr. Rodney Weiher  
Chief Economist  
Officer of Policy and  
Strategic Planning  
National Oceanic and  
Atmospheric Administration

*Prepared by:*

Jeffrey K. Lazo, PhD  
Stratus Consulting Inc.  
PO Box 4059  
Boulder, CO 80306-4059  
(303) 381-8000

December 17, 2002  
SC10160

---

## Executive Summary

The purpose of this report is to provide NOAA with an unbiased expert review of the report entitled “Economic Value of Current and Improved Weather Forecasts in the U.S. Household Sector,” prepared by Jeffrey K. Lazo and Lauraine G. Chestnut of Stratus Consulting, dated May 14, 2002. The Lazo and Chestnut study was undertaken to develop estimates of household values for weather forecast services for use in broad policy analysis and in benefit-cost analysis of weather service programs. Because weather forecast services are generally nonmarket goods, the study used nonmarket valuation (stated preference) methods, which are widely used in environmental economics. Further, because this was the first systematic attempt to estimate these values, and because value estimates from the study would most likely generate widespread interest, NOAA wanted an independent peer review of the methods and results of this work. Some of the major questions addressed by the peer reviewers include whether:

- ▶ the value estimates are based on accepted economic theory and method
- ▶ stated preference approaches are appropriate for the types of values the study intended to elicit and whether they are appropriately implemented and reported
- ▶ the econometric models and methods are theoretically correct and applied correctly
- ▶ there is adequate survey design and development and appropriate survey administration
- ▶ more accurate results would be obtained with a larger more random sample
- ▶ anything in the survey induces specific biases in respondents’ value statements
- ▶ the value estimates are usable for benefit-cost analysis of forecast improvement programs
- ▶ the study method or analysis can be improved.

Stratus Consulting identified and contracted with highly qualified experts in three topic areas to evaluate the reliability and validity of the value estimates. Overall, the reviewers provide a positive evaluation of the methods, application, analysis, and results of the study. The only major issue raised by the reviewers was whether the value estimates are representative of national values. There is a general consensus that where benefits clearly outweigh costs the study provides valid and useful benefit numbers, but for more contentious evaluations, benefit numbers based on a larger and more random national sample would almost certainly provide results readily acceptable for program evaluation and may even provide more information on how people view tradeoffs among weather service products.

---

# 1. Introduction

The purpose of this report is to provide NOAA with an unbiased expert review of various aspects of the report entitled Economic Value of Current and Improved Weather Forecasts in the U.S. Household Sector, prepared by Jeffrey K. Lazo and Lauraine G. Chestnut of Stratus Consulting, dated May 14, 2002. Stratus Consulting identified and contracted with highly qualified experts in three topic areas relevant to evaluating the reliability and validity of the value estimates. This review was undertaken primarily to evaluate the reliability and validity of the value estimates derived from this work. This evaluation was undertaken in context of the goals and objectives of the study and the resources used for this project. Specifically the reviewers were told that “the primary purpose of the review was to assess whether the value estimates derived from this study have ‘usability’ for policy analysis by NOAA and the National Weather Service.”

Overall, the reviewers provide a positive evaluation of the methods, application, analysis, and results of the study. The only major issue raised by the reviewers was whether the value estimates are representative of national values. There was a general consensus that where benefits clearly outweigh costs the study can provide useful benefit numbers. For more contentious evaluations, benefit numbers based on a larger and more random national sample would almost certainly provide results readily acceptable for program evaluation and may even provide more information on how people view tradeoffs among weather service products.

The rest of this report is organized as follows: Section 2 provides information on the three peer reviewers, Professors Bishop, Layton, and Presser; Section 3 describes how the list of issues was developed for the peer reviewers; Section 4 describes the peer review process; Section 5 provides a summary of the peer reviewers’ comments; and Section 6 provides some brief responses and thoughts for potential future work. Attachment A provides the complete “Issues List” described in Section 3, and Attachments B, C, and D provide the peer review comments provided by Professors Bishop, Layton, and Presser, respectively.

## 2. Expert Panel Composition

We sought input from NOAA personnel and the project manager as needed to identify qualified and available external reviewers.<sup>1</sup>

Composition of External Review Panel	
Area of primary expertise	External reviewer
Stated preference (SP) methods	Richard C. Bishop, University of Wisconsin-Madison
Data analysis and econometrics	David F. Layton, University of Washington
Survey implementation and analysis	Stanley Presser, University of Maryland

Brief biographies of each reviewer follow:

**Richard C. Bishop** is Professor and Chair, Department of Agricultural and Applied Economics at the University of Wisconsin-Madison. He is also a member of the faculty of the Institute for Environmental Studies. He joined the Wisconsin faculty in 1973 after completing a PhD and doing postdoctoral research at the University of California-Berkeley. His current duties include departmental administration and research and teaching in environmental and resource economics, with emphasis on welfare theory and benefit-cost analysis, valuation of nonmarket environmental services, sustainability, and renewable resource management. His applied research has focused on a variety of topics, including management of sport and commercial fisheries in the Great Lakes and elsewhere, valuation of environmental amenities, natural resource damage assessment, and endangered species policy. Dr. Bishop is a Fellow of the American Agricultural Economics Association and completed a two-year term as President of the Association of Environmental and Resource Economists in 1998.

**David F. Layton** is Assistant Professor of Public Affairs and Adjunct Assistant Professor of Economics at the University of Washington. He received a PhD in economics from the University of Washington, and was a post-doctoral fellow at Stanford University. He was previously an assistant professor in the Department of Environmental Science and Policy at the University of California-Davis, where he was also a member of the Graduate Groups in Ecology and Statistics. He is an expert in the design and econometric analysis of stated preference surveys for valuing environmental programs. His survey work has included valuing conservation programs for the northern spotted owl and salmon populations, beach recreation in southern

---

1. Norman Meade of NOAA's Damage Assessment and Restoration Program was of considerable help throughout the review process, including identifying qualified external peer reviewers.

California, forest loss due to climate change, the costs of electricity outages to commercial firms and residential customers, and surveys of tourists in Zimbabwe and the Galapagos. His work has been supported by the National Science Foundation, National Oceanic and Atmospheric Administration, National Marine Fisheries Service, a variety of state agencies and private firms, and the International Bank for Reconstruction and Development.

**Stanley Presser**, Professor in the Sociology Department and in the Joint Program in Survey Methodology at the University of Maryland, is known for his research on various aspects of survey methods. He holds a PhD in sociology from the University of Michigan. Dr. Presser was formerly the director of the Survey Research Center at the University of Maryland. He is a past president of the American Association for Public Opinion Research, the main professional association in his field, and was editor of *Public Opinion Quarterly*, the leading journal in the field.

### 3. Preliminary Issue Identification

Working with NOAA personnel, Stratus Consulting develop a list of specific issues to be addressed by the review panel. The issues list defined the objectives of the external review to focus the external experts' efforts on issues most relevant to NOAA's use of the information from the work. NOAA personnel reviewed the issues list to determine if it would satisfy the requirement of NOAA for the use of the information in the report. Based on direction from the NOAA project manager, Stratus Consulting revised and finalized the issues list. This was then provided to the external experts along with a copy of the project report.<sup>2</sup> Attachment A provides the issues list given to the external reviewers.

### 4. External Review Process

Approximately one week after reviewers had received the documents, we had phone calls with the reviewers to provide guidance and feedback as needed. The objectives of the review were reiterated and a verbal summary of the project was provided to facilitate the reviewers' initial understanding of the project.

The external experts were given approximately two weeks to review all relevant documents and to provide written comments in response to the specific issues in the issues list.

---

2. Reviewers were told that all interim reports, data sets, survey instruments, or other materials would be provided as necessary for the review. No additional materials were requested by the reviewers.

When we received all the written comments from the reviewers, we held another conference call to review any unanswered questions and to ensure that we understood all of the reviewers' comments. Attachments B, C, and D provide the written comments from Drs. Bishop, Layton, and Presser, respectively. We did not revise the comments in any manner and did only minor reformatting to incorporate these comments into this document.

## 5. Summary of the External Review

The reviewers employ different approaches and focus on different aspects in their reviews, reflecting their specific expertise. Two broad and consistent opinions arise, as summarized below. Also below we provide a brief summary of each individuals' overall evaluation of the research and specific issues that could be addressed.

1. The reviewers generally support the economic methods, survey design, and data analyses methods in the report. Specific items and recommendations for further evaluation or improvement are noted in the attachments. While many of these items are worthy of evaluation, generally we do not expect these items to have a substantive impact on the results.
2. Two reviewers are concerned about the survey implementation of recruiting individuals to central facilities in selected cities. This method, selected for time and cost control, may have important issues in terms of representativeness within the communities selected, and in terms of the variety of locations around the country. The reviewers are mixed on how this may bias the use of the results: one believes the results are reasonable and usable, one believes the results most likely understate values for some locations, and the survey implementation expert believes there are potential substantial sampling biases.

**Dr. Bishop** reviewed the project materials in large part based on an approach for evaluating the reliability and validity of contingent valuation studies developed in his forthcoming book chapter: "Where to From Here" in a volume edited by Patty Champ, Kevin Boyle, and Tom Brown entitled *A Primer on Nonmarket Valuation* (Kluwer, forthcoming).<sup>3</sup> Dr. Bishop evaluates the content validity, construct validity, and criterion validity. Overall, Dr. Bishop's review indicates a positive evaluation of the methods, application, analysis, and results of the study.

Dr. Bishop's primary concern relates to the "unrepresentative nature of the final respondents." Dr. Bishop's evaluation of potential sampling issues concludes that the value estimates derived in the study may represent lower bound estimates of population values as those more likely to

---

3. This chapter is not included in this report.

have high values were also less likely to be sampled (e.g., rural populations and individuals who work outside more and thus are more affected by weather).

**Dr. Layton** closely followed the issues list in providing his evaluation. Dr. Layton provides a summary to his evaluation: “A complementary way I looked at this report was to decide whether I would feel comfortable with decisions being made on my behalf (as a U.S. citizen) using the values derived from this report. After a thorough review I have determined that the value estimates are indeed ‘useable’ and I feel quite comfortable with basing broad policy analyses on them.” Overall, Dr. Layton’s responses to the issues list indicate a positive evaluation of the methods, application, analysis, and results of the study.

As did Dr. Bishop, Dr. Layton indicates that more information would be helpful on the use of the follow-up “do nothing/status-quo” question in the choice question portion of the survey. He also notes that additional sample may be desirable to develop more precise statistical estimates (i.e., smaller standard errors) but such “a difference is unlikely to be policy relevant.” As did Dr. Bishop, Dr. Layton suggests that presenting a mean or median value from the value of current forecast questions may be unreliable because the median value (about \$110/yr/household) is outside the range of costs presented to individuals (up to \$96/yr/household). He suggests that simply stating that 60% of the sample is willing to pay \$96 still carries significant policy weight.

**Dr. Presser** focused on the issues of sampling error and measurement error. He indicates that sampling error is composed largely of noncoverage and nonresponse. With respect to noncoverage error, Dr. Presser identifies several concerns in the sampling method, including a potential problem with using a sample of residences with listed phone numbers. On reviewing Dr. Presser’s comments, we realized that the description in the report of the random digit dialing approach was unclear (this is clarified in Section 6). Dr. Presser’s evaluation of the sample-population comparisons in the report leads him to conclude that there may be significant differences in values between the population and the sample, although it is not possible to determine a priori in what direction such differences may occur.

With respect to measurement error, Dr. Presser states that “the report demonstrates commendable concern for the effects of measurement error.” He then identifies some concerns about survey wording and information content. He does not discuss how these may have affected value estimates. These issues could be addressed in future implementations of the survey instrument.

## 6. Response and Potential Future Work

In undertaking the central site implementation, we were aware that this would not be the same as a random national. Overall the variability of values across the different locations and individuals is expected and is found to be reasonable with expectations. We chose cities that cover the broad range of climatic conditions individuals face across the country, from relatively benign weather in San Diego to highly variable and less predictable weather in Billings, Denver, and Albany. Overall results from the study suggest that values for weather forecasts are more a function of weather variability and the individual's use of forecasts than they are of sociodemographic characteristics such as income or education. Thus the impact of having a sample with higher average education and income than their cities' populations may have minimal impact on the representativeness of the value estimates.

After reading Dr. Presser's comments, we reviewed the sampling methods with the subcontractor that implemented the survey (PA Consulting Group). They provided additional information ensuring that the random digit dialing (RDD) portion of the survey was truly RDD. Specifically it has been clarified that Survey Sampling, a subcontractor to PA, provided a 10-digit randomly generated number. The samples were drawn from a radius around the central site based on a 7-mile radius (Portland; San Diego; Columbus; Oklahoma City; Billings; Clifton Park, New York; Madison) or a 10-mile radius (Miami, Denver). Numbers were eligible for the sample if any part of the zip code they were in fell within the 7 or 10 mile radius. For the residential sample, Survey Sampling then screened out business, fax, and nonworking numbers. Thus, everyone with a phone in these zip codes had an equal probability of being selected regardless of whether or not their phone is listed.

Stratus Consulting will undertake minor editorial revisions to the final project report in response to comments from the reviewers. These include clarifying the affiliation of Dr. Tourangeau and including a more accurate explanation of the random digit dialing method used for subject recruitment as described above. Based on additional comments, we will also revise the current executive summary to include the core value estimates derived in the study.

Pending funding, future work could include the following:

1. Additional econometric modeling and analysis: Drs. Bishop and Layton provide suggestions for improving the evaluation of responses by undertaking additional or alternative econometric analysis. While they do not feel that the analysis would substantively change the results, such additional analysis would lend further interesting and worthwhile understanding of the respondents' perceptions of and values for current and improved weather forecasts. It most likely won't address the issues of limited sample using the central site implementation.

2. Survey revisions: As expected, all three reviewers noted a number of places in the survey where they feel clarification of language may improve the instrument. While some of these had been addressed in pre-testing and not thoroughly documented in the current project report, we would explore further revision and improvement in the survey instrument based on these suggestions. One potential change would be to again increase the upper bound indicated to individuals for the cost of current forecasts to better define the shape of the demand curve and to better identify statistically acceptable mean and medium values.
3. Implement a revised survey instrument with a national random sample: Drs. Bishop and Presser indicate that addressing potential sampling issues may be best achieved by implementing the survey again with a more random national sample. Dr. Layton suggests that doing so would most likely not provide substantively different value estimates. Stratus Consulting will provide a separate memorandum indicating the potential costs for implementing such a national random sample with subsequent analysis and reporting.

It should be noted that undertaking items 1 and 2 above will not entirely address the concern of the reviewers regarding sample selection, although additional data analysis using appropriate population weights may adjust for some potential sample biases.

## ATTACHMENT A: ISSUES LIST

### *Economics*

- ▶ Are the valuation method and analytical technique(s) used in this study appropriate for the task?
- ▶ Are the value estimates derived from this study consistent with accepted economic theory and method?
- ▶ Are the value estimates derived from the data analysis and econometric analysis correctly presented?
- ▶ Are there other economic issues that have not been addressed in the study or in the report?

### *Stated Preference Methods*

- ▶ Is the use of stated value and stated choice approaches appropriate for elicitation of the types of values this study is intended to elicit? Are there other feasible or appropriate methods?
- ▶ Have the stated value and stated choice approaches been appropriately implemented and reported?
- ▶ Have the choice sets and different versions of the survey been appropriately designed?
- ▶ Are there specific biases in the application of SP methods that have not been addressed which would significantly affect the basic results of the research?
- ▶ Are there other stated preference issues that have not been addressed in the study or in the report?

### *Data Analysis and Econometrics*

- ▶ Is there adequate data quality control-quality assurance?
- ▶ Is the basic data analysis correct and adequate?
- ▶ Are the econometric models and methods theoretically correct and applied correctly?

- ▶ Can weighting of the existing data set with appropriate socioeconomic data improve the accuracy and applicability (i.e., to a broader cross section of the population) of the results?
- ▶ Are there other data analysis or econometric issues that have not been addressed in the study or in the report?

### ***Survey Methods***

- ▶ Was there adequate survey design and development (including pre-testing, revision, and review)?
- ▶ Is the quality of the survey instrument sufficient to elicit the desired information from respondents (e.g., how appropriate was the questionnaire design/wording)?
- ▶ Is the information content in the survey appropriate and does there appear to be any excessive cognitive burden on respondents?
- ▶ How appropriate was the sample selection process and size?
- ▶ How appropriate was the survey administration?
- ▶ How much more accurate would the results have been had a larger and more random sample been used?
- ▶ Is there anything in the survey design or implementation that would induce specific biases in respondents' value statements?
- ▶ Are there other survey design or implementation issues that have not been addressed in the study or in the report?

### ***Overall***

- ▶ Have the researchers left out any important ideas/techniques from the relevant empirical literature?
- ▶ Are the basic results/conclusions of the study reliable, i.e., capable of being replicated by an independent third party?
- ▶ Are the value estimates usable for benefit-cost analysis of weather service improvement programs that would improve weather forecast accuracy in manners similar to those described in the survey instrument?

*Additional Concerns or Suggestions*

- ▶ What related questions (other than the ones directly addressed by the study) could or should be addressed with the existing data set?
- ▶ How could the overall study (methodology, analysis, interpretation of results, conclusions, etc.) be improved?

**ATTACHMENT B: COMMENTS PROVIDED BY RICHARD C. BISHOP****Review of “Economic Value of Current and Improved Weather Forecasts in the US Household Sector”**

**Reviewer: Richard C. Bishop, Professor and Chair, Department of Agricultural and Applied Economics, University of Wisconsin, Madison, WI 53706, (608) 262-8966, [bishop@aae.wisc.edu](mailto:bishop@aae.wisc.edu)**

**Date: October 14, 2002**

The instructions wisely call for a systematic approach to the review, but allow reviewers to use their own framework, which I will do. The basic question here is whether the results of this study are sufficiently valid to be used in policy analysis. Hence, I will organize my remarks around the “Three Cs” of validity assessment, content validity, construct validity, and criterion validity. My latest thinking on the Three Cs as they apply to valuation studies will appear as the final chapter of a volume edited by Patty Champ, Kevin Boyle, and Tom Brown entitled, *A Primer on Nonmarket Valuation* (Kluwer, forthcoming). Attaching a copy to this review will allow me to avoid having to repeat long explanations of where I am coming from in this review.

The goal of valuation is to measure the true values of the items in question, here either the value to households of the weather forecasts they currently receive or the values they would receive from improved forecasts. What makes this tricky is that true values are unobservable (attachment, pp. 3-4). We can’t see inside people’s heads to observe true willingness to pay (WTP). We can only look either at behavioral evidence (e.g., behavior in the marketplace) or their responses in surveys to stated-preference questions. In the case of weather forecasts, there does not appear to be sufficient behavioral evidence to arrive at values, so stated-preference measures must be used.

Content validity has to do with procedures. To what extent were the procedures used throughout a study conducive to measuring true values? Construct validity is usually broken down into two parts, theoretical validity and convergent validity. In theoretical validity testing, study results are tested against prior expectations, which are in turn based on theory and intuition. Convergent validity involves comparing results of two different ways of measuring something. Criterion validity assessment occurs when a proposed measure of something is compared to a measure that is widely accepted as valid. All this is explained in more detail in the attachment, pp. 6-8.

Criterion validity assessment is most relevant when thinking about the validity of valuation methods and approaches. We basically ask whether a method or approach is capable of producing valid values in a specific application if the study in question was done well (has high content validity) and tests out well in construct validity assessment. So, it is relevant in the current context to ask whether contingent valuation — what this study calls the “stated-value” or SV approach — is a valid approach. As the report points out and as I argue in more detail in the attachment (pp. 15-19), there is a growing body of evidence to support the conclusion that contingent valuation studies — if done well — can provide values that are sufficiently valid to be used in policy analysis. We know less about the stated-choice approach because it is newer, but so far the evidence is positive. Much that has been learned about contingent valuation must apply to stated choice as well. In fact, it would be easy to think of stated-choice questions as simply another form of contingent valuation. Furthermore, my impression is that stated-choice studies have generally stood up well in theoretical validity testing. Thus, judgments about the validity of the study before us must rest on whether it was done well and whether it tested out well in terms of construct validity.

### *Content Validity*

To look systematically at content validity, I have proposed a set of twelve questions that can be asked. These are explained in the attachment and the reader who wants to learn more about the rationale for the questions can look there. I will simply state each question here and address each in turn. Because the attachment was designed to address environmental valuation, I have made minor modifications in the questions to adapt them to study under review.

1. *Was the true value clearly and correctly defined?* A valid study begins with a clear vision of what is to be measured. The Stratus study rests on a sound theoretical foundation. At the center is the economic theory of the value of information, which is appropriate. And I liked the richness that the report adds to the bare bones theory with Figure 2-1 and associated text.
2. *Were the respondent-relevant attributes fully identified?* Theory only goes part way toward identifying what is to be valued. Any valid valuation study needs to clearly define which attributes of the items to be valued actually do matter to potential study subjects. Through consultation with NOAA people, past literature, and focus groups, this study did a good job in this regard.
3. *Were the potential effects of the intervention adequately documented and communicated?* “Intervention” here refers to the steps taken to improve weather forecasts. The study seems to have done a good job of figuring out what sorts of improvements in weather forecasts might be made and in communicating them to respondents. In particular, on the communications side, I have found one-on-one interviews of the type described in

chapter 3 of the report to be a very effective tool for working the bugs out of stated-preference questions. Had there been problems in communicating the effects of the interventions, they would hopefully have turned up there and been rectified.

4. *Were the respondents aware of their budget constraints and of the existence and status of substitutes?* Question 14 served to remind people of their budget constraints. I would not have done it that way, but the way it was done achieved the desired end and probably didn't do any harm. For the choice questions, the substitute is rather clearly the real world status quo, which was explained in the survey instrument between Questions 14 and 15. So I think the issues raised by this question were adequately addressed.
5. *Was the context of valuation clearly specified and incentive compatible?* The context for valuation refers to the various terms of the proposed transactions in the valuation questions, not only what services will be provided under Alternatives A and B or under Question 33, but also when they will be provided and how they will be paid for. Most respondents are familiar the NWS and understand that it is supported through taxes, and the instrument clearly points out that improvements will be paid for through taxes. The survey instrument does not appear to specify the timing of either the new taxes or the forecast improvements. This is a technical flaw, but it is probably minor in this case. Left to their own devices, respondents probably thought they would have to pay starting soon and that the improvements would be up and running in a few years. The issue of incentive compatibility is an interesting one. Other things being equal, one would not want to design valuation questions with built-in incentives to respond in ways that do not reflect underlying preferences. I do not recall having seen this addressed for SC questions. Intuitively, the form used in this study looks incentive compatible to me, but it would be nice to see a formal treatment of the issue in the literature. Adding comparisons to the status quo (e.g., Question 16, 18, etc.) after each pair of alternative is a little worrisome in this regard, but I can't intuitively visualize an incentive compatibility problem there. Question 33 uses a conventional contingent valuation payment card mechanism. If respondents treat this like they would a first-price auction, there could be an incentive issue here, but it probably is not a big one. If there is a bias, it would be downward.
6. *Did survey respondents accept the scenario? Did they believe the scenario?* Some scenario rejection is always a problem in these studies. If anything, I suspect this study was less prone to it than a lot of others I have seen. Furthermore, the study introduces a very innovative way to test whether scenario rejection is having an effect. This involved using factor analysis on responses to Question 34 and using the results as independent variables in the valuation equation for Question 33. Some scenario rejection was indicated. Unless I've missed it elsewhere in the literature, this procedure is original to this study. If so, I hope the authors will publish it. Two questions do arise in this regard, though. First, if the

effects of scenario rejection have been identified statistically, why not explore the implications of correcting for it? If there are to be revisions of this report, I would recommend that this issue be explored. Second, if scenario rejection affected the SV question, might it also have affected the SC questions? This issue is definitely on the frontier of SC methodology and probably cannot be answered at this time but it would have been good to acknowledge it.

7. *How adequate and complete were survey questions other than those designed to elicit values?* The study does quite a lot in this regard, supporting the valuation work both directly with potential variables to use in the analysis of values (e.g., income, Question 4, and questions on work and leisure activities outdoors) and indirectly with variables to use in support of the valuation results (e.g., the importance, satisfaction, and adequacy-of-forecast-attributes questions). I could be picky about the way some of the questions were designed, but what researchers did was probably up to the task.
8. *Was the survey mode appropriate?* The survey instrument was simple and straightforward enough to be self-administered (i.e., interviews were not necessary for the final survey). Using market research facilities is relatively unprecedented in this type of study, and I am not convinced that it was cheaper than a mail survey with say 1000 subjects, but I could be mistaken. Doing a self-administered survey at market research facilities in selected cities did have one drawback. It committed the team to a sample that was not very representative of the population. More on this under question 10 below.
9. *Were the qualitative research procedures, pretests, and pilots sufficient to find and remedy identifiable flaws in the survey instrument and associated materials?* The short answer is probably “yes.” The focus groups, one-on-one interviews, and pilot provided mechanisms to be fairly thorough in this regard. Conducting only 11 one-on-ones seems a little thin on sample size, but should have been enough to weed out the worst of the problems. The report was not as informative as it might have been in summarizing the results of the focus groups and one-on-ones. Section 3.1.2 describes what was done, but says little about what was learned. The researchers sort of ask us to assume that in doing what they did, they did what needed to be done. Knowing them, I would tend to believe them but would have liked more concrete information. The next section of the report, describing the pilot, does better in this regard.
10. *Given the study objectives, how adequate were the procedures employed to choose study subjects, assign them to treatments (if applicable), and encouraging high response rates?* Most readers will not be surprised when I say that I think this is the studies largest weakness. Relative to the geographic distribution of the national population, the sample was highly skewed against the East and probably the South. And the results show that geography matters. At the next level, few could argue with a straight face that the cities

are representative of their regional populations. And, adjusted samples in the hundreds or even thousands for each city dwindled steadily until only 30 to 40 showed up to do the survey. Table 4-3 shows clearly (and honestly) that the respondents were far more highly educated than the populations in their cities and that is surely only the tip of the iceberg. That the investigators could be forced by budgets into such shortcuts is understandable. Their conclusion (p. 6-3) that their results “provide unbiased, reliable, and valid value estimates” seems like a bit of a stretch. This flaw will cloud the use of their results in future policy analyses. I will talk about what might be done to address this problem in the “Conclusions” section.

11. *Was the econometric analysis adequate?* The econometric analysis that was done seems fine, but at this stage it looks underdeveloped to me, particularly with respect to the SC data (Section 5.2). The analysis there seems unusually sparse given the state-of-the-art in stated-choice research. An appendix lays out some foundations for a much deeper analysis that would have accounted for heterogeneous preferences in an original way. That these ideas were not implemented may be another symptom of a very tight budget at the end of the study. More analysis of the SC data might have yielded additional insights about the theoretical validity of the values for improved weather services. Also, I will suggest below that an SC model with heterogeneous preferences might be used in a convergent validity test with a parallel SV model.
12. *How adequate were the written materials from the study?* I have few complaints about the report. One can easily tell what was done and how and what the results were. I sometimes wished for a bit more documentation of some of the arguments. For example, they suggest (p. 3-6) that the extra stuff in the right-hand column of Question 15 “has been shown to enhance respondents’ ability to understand the choice process.” Where? However, when the budget started to bight, there were probably higher priorities than thorough documentation of sources. It might also have been useful to explore the policy implications of the SC results in a bit more depth. Tables 5-4 and 5-5 focus on monetary values. While valuation was a primary motivation for the study, all the fuss and bother of SC questions was needed only if it was also desirable to consider the tradeoffs between different types of forecast improvements. It might be useful to NOAA in planning future efforts to improve forecasting to know the marginal rates of substitution.

I want to comment on two other aspects that have not yet come out. One is the design of choice pairs for SC questions. I do not know a lot about this subject, but as near as I can tell, what they did is right on. Stratus seems to be ahead of the curve in doing this sort of thing, which probably paid off in this study.

Second, there seems to be wide agreement among those working in the field that embedding can be a problem in studies like this, but I don't think it was diagnosed or dealt with very effectively using Question 35. I think — and admittedly others might disagree — that such questions invite respondents to wander off into speculations about embedding that are not very meaningful. And to tell them that they might have had difficulty disentangling their unembedded value, and then ask them what percentage it is, seems strange to me. If I thought, based on focus groups and one-on-ones, that embedding could be a serious problem, I would rather try to head it off by explaining the problem to respondents just before the contingent valuation question and asking them to do their best to value only the item as described. Also, the report does not acknowledge or try to address the possibility that, if embedding is a problem with the SV question, why not with the SC questions? I would strongly recommend that NOAA avoid using the values from question 33 that are adjusted in this way for embedding.

By way of summarizing my content validity assessment, this study seems to have produced a strong survey instrument with a good theoretical foundation. It also deserves good marks for reporting. Unfortunately, the way the study was implemented leaves it open to concerns whether the results are representative of the US population and the direction or potential magnitude of the bias is not addressed in the study. And more could be done on the econometric side.

### *Construct Validity Assessment*

Construct validity assessment is explained in the attachment, pp. 11-12. Consider first the theoretical validity of the SC results. Table 5-3 contains strong evidence that respondents did consider the tradeoffs between the different possible improvements in forecasting services and between those services and money. This is quite encouraging since it coincides with how economists expect (or hope) people will think about these types of questions. Additional analysis using models that account for heterogeneous preferences might add additional insights about the strength of the results.

By the way, I would not worry much about the negative coefficient on “frequency.” Particularly given other results of the survey, it seems very plausible to me that many respondents found that increased frequency detracted from the appeal of an alternative. If members of the species *Homo economicus* have no interest in increasing frequency, then as the authors put it (p. 5-4), they “could simply ignore the updates and be no worse off.” But most of us can envision *Homo sapiens* saying, “I am less interested in an alternative where you would spend my money to increase frequency.” Just conclude that frequency has no value and get on with it, as was done at the bottom of Table 5-5.

Results of the theoretical validity testing of the SV results in Table 5-9 produced mixed results. I am not sure what I think of the models where WTP was weighted by responses to Questions 38 and 39. This is unfortunate since the “weighted WTP (raw)” model has the most positive results.

Perhaps the authors can address this procedure in the upcoming conference call. The other models have many variables that are insignificant or only marginally significant. Furthermore, there does not seem to be very much robustness in what is and is not significant across the four models. And the poor performance of “multiday” in all the models is particularly disappointing given its strong showing in the analysis of the SC data. Respondents seem to have struggled some with the SV question (Question 33), which introduced a lot of noise into the data. This seems odd given the similarities between the SC questions and the strong preliminary results for the SC questions as just noted. Perhaps if the data were somewhat noisy, this combined with only one observation per respondent may have made the sample size too small to get more significant variables. Perhaps additional econometric analysis would help. I will also suggest some work on convergent validity in a moment. In the meantime, I am somewhat uneasy about the validity of the SV values.

Table 5-15 contains evidence that Question O2 (which addresses the value of current weather information) worked reasonably well. A good strong coefficient on “current cost” was good to see. Education/income is significant as is weather variability. The “discretionary use factor” and “hours spent traveling to work” are marginally so. The usefulness of the results in estimating value is limited by the fact that the “current-cost” amounts used in the questions did not go nearly high enough to gauge where the right-hand tail of the distribution of values lies, but perhaps the policy relevance of such a mean dollar estimate would be limited given that such an average would be far above per household cost. That current weather services are economically justified is evident from the data as it now stands.

Turning to convergent validity, there may be an unexploited opportunity for a test here using the SC and SV data together. We can already see hints of convergent validity in the roughly comparable SC and SV values for maximum improvements and in the strong performance of “one day” and “geographic” in both Tables 5-5 and 5-9. The basis for a full test of convergent validity is found in the very close correspondence between the SC and SV questions. If comparable heterogeneous-preference models could be constructed and estimated for both, the null hypothesis would be that responses to the SC and SV questions drew on the same underlying utility functions. If the null is accepted, this would allow results from the SC and the SV questions to be mutually supportive and would go some way toward reassurance that the SV question “worked.” It might even be possible to simply combine the two parts of the data in a single grand model. If the null is rejected, then econometric criteria could be applied to judge which data are more valid. All this would require considerably more econometric muscle than I possess, but it seems to me to be doable in principle.

## *Conclusions*

So, we started out asking whether the results of this study are sufficiently reliable to be used in general policy discussions and benefit-cost analysis. As I have pointed out, I think it has a number of strengths, particularly for the first study in this area. It developed a survey instrument with skill and care, doing the upfront qualitative stuff and a pilot test to work out most of the bugs. The as yet preliminary analysis indicates that careful attention to the design and execution of the SC exercise probably yielded strong results. The method of gauging scenario rejection seems very innovative and potentially useful. On the minus side, my biggest concern is with the unrepresentative nature of the final respondents. I also see evidence that perhaps Question 33 did not work as well as the investigators had hoped. I would like to have seen more econometrics along the lines that I have suggested in order to enlarge the basis for construct validity assessment. My guess is that such work would support the validity of the results at least so far as the SC part of the study is concerned. But such positive conclusions — if they are indeed forthcoming — would not mask the study's flaws relating to the unrepresentativeness of the sample.

Thinking about future policy discussions and benefit-cost analyses, I find the kind of discussion in “A Thought Experiment” (Section 6.2) plausible up to a point. NOAA could do that sort of policy analysis using the results of this study as they now stand. However, unless benefits exceed costs by a wide margin, the conclusions from such analyses will be vulnerable to attack. Three directions for shoring up the work can be suggested.

1. *Do more econometrics.* I doubt that this would be enough to rectify the sampling problem. It would be quite a jump to adjust the results using socio-demographic characteristics of the population. Maybe that could be done, but I doubt it. What more econometrics could do is strengthen the case for the construct validity of the SC results and perhaps strengthen the case for the SV results for the existing sample.
2. *Explore the case for calling the SC and possibly SV values theoretical lower bounds on the population true values.* Some thought would have to be put into this, but let me sketch how the argument might go. The current sample would seem to be overly balanced toward the western and perhaps central U.S. Could one argue that the more populated areas of the West, the coastal areas, have less severe and more predictable weather than in the East and South? If so then, all else equal, would it follow that improved forecasts would be more valuable to the underrepresented East? This would make the values in the report underestimates. Also, better-educated people are over-represented. People with less education tend to be the ones that work outside and hence have a higher value for weather information. (Unfortunately, better-educated people also have higher incomes and the separate effects of education and income have yet to be teased out in the analysis. At the risk of being a “Monday-morning quarterback” (with 20-20 hindsight), it is

unfortunate that the survey did not ask for occupation, with an emphasis on outdoor jobs. The same might be said for leisure activities. Perhaps this is the sort of thing one learns from the first study of it kind.) The sample probably has some bias in favor of urbanites and rural people may be more likely to work and engage in leisure activities outdoors. If such an argument could be constructed, it would help offset criticisms based on sampling.

3. *Do a national survey.* I think the survey instrument is very close to being strong enough to be administered by mail. A sample of 1000 or 1500 ought to be enough to get estimates for the population with fairly narrow confidence intervals. Such a survey would do much to increase confidence about the values Americans place on weather forecasting services.

## **ATTACHMENT C: COMMENTS PROVIDED BY DAVID F. LAYTON**

### **Peer Review of the “Economic Value of Current and Improved Weather Forecasts in the U.S. Household Sector.”**

**Peer Reviewer: David F. Layton, Ph.D.**

**Summary:** I was asked to review the “Economic Value of Current and Improved Weather Forecasts in the U.S. Household Sector” to assess whether the value estimates derived from this study have “usability” for policy analysis by NOAA and the National Weather Service. A complementary way I looked at this report was to decide whether I would feel comfortable with decisions being made on my behalf (as a U.S. citizen) using the values derived from this report. After a thorough review I have determined that the value estimates are indeed “useable” and I feel quite comfortable with basing broad policy analyses on them.

Below I address the specific questions suggested as part of this review. In places I discuss at some length, and in other places I respond simply with “Yes” when I do not have much to add.

#### **6.1.1 Economics**

- ▶ Are the valuation method and analytical technique(s) used in this study appropriate for the task?  
**YES**
- ▶ Are the value estimates derived from this study consistent with accepted economic theory and method?  
**YES**
- ▶ Are the value estimates derived from the data analysis and econometric analysis correctly presented?  
**YES**
- ▶ Are there other economic issues that have not been addressed in the study or in the report?

**There does not appear to be any important omissions.**

**Stated Preference Methods**

- ▶ Is the use of stated value and stated choice approaches appropriate for elicitation of the types of values this study is intended to elicit? Are there other feasible or appropriate methods?

**The use of stated value/choice questions is appropriate for these type of values — values which are at best extremely difficult to get at with revealed preference methods (and probably impossible). Other feasible methods would, I believe, all be within the realm of Stated Preference methods. That is, one could use other elicitation tasks, but it would appear to me that some survey based approach is necessary.**

- ▶ Have the stated value and stated choice approaches been appropriately implemented and reported?

**YES**

- ▶ Have the choice sets and different versions of the survey been appropriately designed?

**YES, clearly great care was taken in the generation of the choice sets, both from a statistical standpoint and from an information and respondent reasonableness standpoint.**

- ▶ Are there specific biases in the application of SP methods that have not been addressed which would significantly affect the basic results of the research?

**I do not believe so.**

- ▶ Are there other stated preference issues that have not been addressed in the study or in the report?

**I do not see any significant omissions.**

**6.1.2 Data Analysis and Econometrics**

- ▶ Is there adequate data quality control-quality assurance?

**YES**

- ▶ Is the basic data analysis correct and adequate?

**Yes, the data analysis is correct and adequate.**

- ▶ Are the econometric models and methods theoretically correct and applied correctly?

**YES, Appendix F is quite clear, correct, and impressive.**

- ▶ Can weighting of the existing data set with appropriate socioeconomic data improve the accuracy and applicability (i.e., to a broader cross section of the population) of the results?

**Using socioeconomic data to re-weight is justified given that the random sampling is city specific (which is not random), so an exploration of reweighing is appropriate.**

- ▶ Are there other data analysis or econometric issues that have not been addressed in the study or in the report?

**I see no crucial issues that have not been addressed, but mention some topics that are worthy of consideration at the end of the report.**

### 6.1.3 Survey Methods

- ▶ Was there adequate survey design and development (including pre-testing, revision, and review)?

**The survey design and pre-testing are very careful, thorough, and appropriate. The types of attributes and reasonable ranges for them were based on work by scientists. Then great care was taken in translating the scientific/technical details to meaningful descriptions of attributes. The researchers used the full range of pre-testing approaches including focus groups, verbal protocols, and significant expert review.**

- ▶ Is the quality of the survey instrument sufficient to elicit the desired information from respondents (e.g., how appropriate was the questionnaire design/wording)?

**Yes, this is a high-quality stated preference instrument. My comments regarding survey wording are relatively minor. These are:**

**The graphics under Q 11 could be labeled more clearly. In Q 12, graphic detail of the kind used in Q11 with shading for area sizes would have been helpful.**

**Under “What Program Do You Prefer” the second bullet mentions presenting a range of costs, but costs are not actually presented as range, but instead as just one number (say \$3 per year more).**

- ▶ Is the information content in the survey appropriate and does there appear to be any excessive cognitive burden on respondents?

**Q10 seems is written in such a way as to not be easy to process, but in any event does not appear to be crucial to the overall study results.**

The number of stated choice questions in conjunction with the other questions does not seem “light.” It would have been nice to have more discussion of how the pre-testing/review process ensured that the cognitive burden was not excessive. But based on the very careful review process, and on my own experience with repeated choice scenarios, I am not greatly concerned that this survey posed an excessive cognitive burden.

- ▶ How appropriate was the sample selection process and size?

The sample size is reasonable, especially given the repeated choices. I would have liked to have seen a complementary set of cities surveyed from disparate areas of the nine regions. For instance, Miami allows one to capture a particular kind of weather experience and values, but it is at the extreme southern end of the southeast which includes Richmond, Virginia which has different weather than Miami. While the report did not pursue regional models (quite reasonably give the resource constraints), pairs of cities in each region would help one to better explore the impact of different types of weather/climate on weather forecasting values. This would be policy relevant if the agency were interested in offering differing mixes of improved services across regions.

- ▶ How appropriate was the survey administration?

The survey administration using self-administered surveys is very appropriate for this kind of instrument.

- ▶ How much more accurate would the results have been had a larger and more random sample been used?

Given the overall high significance level for the important models, additional sample from the same cities would not change the accuracy or precision much. Obviously more sample will result in lower standard errors, other things being equal, but given the overall significance level it will take a great deal of additional sample to make much of a difference, a difference which is unlikely to be policy relevant. On the other hand, having an additional sample from nine other cities would improve the coverage of weather areas and would be expected to improve the robustness of the results (as opposed to just reducing standard errors), and would be helpful if interested in addressing the impacts of weather variation on values.

- ▶ Is there anything in the survey design or implementation that would induce specific biases in respondents’ value statements?

I did not pick up on anything that would lead to biases.

- ▶ Are there other survey design or implementation issues that have not been addressed in the study or in the report?

**As I noted earlier, it would be helpful to discuss the examination/analysis of choice task complexity that occurred as part of the pre-testing/design phase.**

#### 6.1.4 Overall

- ▶ Have the researchers left out any important ideas/techniques from the relevant empirical literature?

**It would be helpful for the authors to offer an assessment from the literature (published peer reviewed, or from working paper/report literature) of the equivalence or comparative advantage/disadvantage of the dichotomous choice + secondary status quo stated choice questions versus a straight three alternative choice. It would also be helpful if they could supplement this with information from the pre-test design phase.**

- ▶ Are the basic results/conclusions of the study reliable, i.e., capable of being replicated by an independent third party?

**YES, I believe that the results are replicable both in terms of someone else using their data, or in terms of generating a comparable instrument and collecting new data.**

- ▶ Are the value estimates usable for benefit-cost analysis of weather service improvement programs that would improve weather forecast accuracy in manners similar to those described in the survey instrument?

**YES.**

#### 6.1.5 Additional Concerns or Suggestions

How could the overall study (methodology, analysis, interpretation of results, conclusions, etc.) be improved?

**The following are topics that could be considered.**

**The authors should consider expanding the random parameters model in Appendix F to include more random parameters given that the single random parameter models clearly fit better. They should see how the values from these models compare to the models without random parameters.**

The authors could explore whether the value of some attributes might be better modeled using non-linear transformations. I would view this as fine-tuning and does not affect my overall assessment of the research as being useful and useable.

I would like to see models without the factors scores as well, for a number of reasons. First, there is no correction in the standard errors for the fact that factor scores are estimated quantities. Second, I would rather see a judicious choice of variables/responses entered directly into the indirect utility function as opposed to pre-processing via the factor analysis. This is an area of evolving research — and I doubt that this dramatically changes the results.

I would suggest that in the analysis of the current value of forecasts the authors should either focus on the within bid range information (that more than 60% of the sample is willing to pay at least \$96) and not the median, or if obtaining a prediction of the median outside of the bid range is crucial, then they should consider a flexible distributional form for estimation (for instance a generalized gamma distribution). We can expect some sensitivity of the median to distributional assumptions given that the median is more than the highest bid, and in this case starting with a flexible form is justified.

## **ATTACHMENT D: COMMENTS PROVIDED BY STANLEY PRESSER**

Evaluation of the May 14, 2002 report “Economic Value of Current and Improved Weather Forecasts in the U.S. Household Sector” by Jeffrey K. Lazo and Lauraine Chestnut

by Stanley Presser

In many respects, this is a carefully done study that should serve as a very useful resource for all future work on estimating the value of weather forecasting. However, I have a few concerns about the study’s estimates of willingness-to pay.

All survey estimates contain error; some due to measurements that are not made (sampling errors), and others due to the way measurements are made (measurement errors). In the Lazo-Chestnut (hereafter LC) study the errors from these sources may be very large.

### **SAMPLING ERRORS**

The two most troublesome kinds of sampling error stem from noncoverage (some members of the population are missing on the list from which the sample is selected) and nonresponse (some members of the sample that is selected are not measured).

Noncoverage: In the LC study, the sample was limited to household telephone numbers that Survey Sampling Inc., identified as being located within seven miles of a particular focus group facility. The report does not indicate how this identification was made. Geographical information is usually available only for households with listed phone numbers. Thus households with unpublished phone numbers (which account for a large fraction of all households) may have had no chance of selection. In addition, the roughly five percent of households without a telephone could not be sampled. Finally, households more than seven miles from the focus group facilities, which account for an unknown but potentially large portion of the population, apparently had zero chance of selection. Households with listed phone numbers differ in various ways from those with unpublished listings; households with phone service are very different from those without service; and households less than seven miles from the chosen facility may have been substantially different from those further away from the facility (for example, depending on the facility’s location, they would have been disproportionately either more or less inner city).

Nonresponse: Ideally one wants to make measurements on 100 percent of the eligible members of the selected sample. In practice this is rarely possible, as some members of the sample refuse and others are not located. In the LC study, the response rate (the proportion of the eligible sample on which measurements were made) was extremely low: less than 5 percent (the rates in Table 3-7 multiplied by those in Table 3-8). People who are easier to contact and willing to

cooperate with survey requests tend to be different in diverse respects from harder to reach and less cooperative individuals.

The key question is therefore: Did noncoverage and nonresponse in the LC study bias its willingness-to-pay estimates? Bias is a multiplicative function of the amount of noncoverage (or nonresponse) and the distinctiveness of the noncovered (or nonresponse) households. Because the proportions of noncovered and nonrespondent households were so large, even modest differences in willingness-to-pay between noncovered/nonresponse households and those that were interviewed would lead to nontrivial bias in the estimate of value for weather forecasts.

Although it is difficult to know with certainty whether noncovered or nonresponse households had distinctive values for forecasts, the sociodemographic composition of the sample that was interviewed provides an indicator of the potential magnitude of the problem. According to Table 4-3, the survey respondents were very different from the general population. For instance, whereas only about half the adult population in all nine cities were high school graduates, almost all the survey respondents had graduated from high school. And whereas Miami was the city with the largest elderly population (21 percent were 65 years or older), the Miami sample had the smallest elderly proportion (only 2 percent). Since there is good reason to believe that characteristics like age and education are related to the value placed on forecasts, these sampling problems may be a major source of bias in the survey's estimate of forecasting's value.

## MEASUREMENT ERRORS

The report demonstrates commendable concern for the effects of measurement error. For instance, the inclusion of Question 35 allowed the authors to discover that most respondents reported a value for forecasting improvements that was not restricted to the normal weather conditions that were supposed to be the exclusive focus of the respondents' consideration. (The reference on page 5-8 to Question 33, as opposed to 35, is presumably a typographical error.) The authors addressed this problem by using the answers to Question 36 to adjust the willingness-to-pay estimates. Yet those answers may themselves contain substantial error, as the task posed by Question 36 is a very demanding one, and we know that many adults do not fully understand the idea of percentages. Moreover, even though respondents had probably never done a task like that posed by Question 36, they were given little guidance in how to do it (and they were provided no information about whether severe weather forecasts could be improved without also improving those for normal weather).

Fewer respondents would have included severe weather forecasts in their willingness-to-pay answers if the questionnaire had emphasized that forecasting of only normal weather was the service to be valued. The instruction at the beginning of the questionnaire, "in this booklet we are talking mainly about normal weather conditions," was ambiguous and it was immediately undercut in the very first question's reference to "warnings." More frequent and clearer

references throughout the questionnaire to the normal weather restriction (especially as part of the wording of Question 33) would have largely eliminated the problem.

On a related issue, respondents may not have received sufficient information about some of the improvements they were asked to value. For instance, it is not clear whether informed decisions about the utility of greater geographical detail can be made without knowing how often (and by how much) weather varies within the present 30 by 30 miles resolution. Yet that information was not provided. Thus, to the extent that respondents' assumptions about the matter were incorrect, their valuations could have been distorted.